MULTILINGUAL RESOURCES FOR NLP

IN THE LEXICAL MARKUP FRAMEWORK (LMF)

Gil Francopoulo[1], Nuria Bel[2], Monte George[3], Nicoletta Calzolari[4], Monica Monachini[5], Mandy Pet[6], Claudia Soria[7]
[1]TAGMATICA: gil.francopoulo@wanadoo.fr
[2]Universitat Pompeu Fabra: nuria.bel@upf.edu
[3]ANSI: dracalpha@earthlink.net
[4]CNR-ILC: glottolo@ilc.cnr.it
[5]CNR-ILC: monica.monachini@ilc.cnr.it
[6]MITRE: mpet@mitre.org
[7]CNR-ILC: claudia.soria@ilc.cnr.it

ABSTRACT: Optimizing the production, maintenance and extension of lexical resources is one the crucial aspects impacting Natural Language Processing (NLP). A second aspect involves optimizing the process leading to their integration in applications. With this respect, we believe that a consensual specification on monolingual, bilingual and multilingual lexicons can be a useful aid for the various NLP actors. Within ISO, one purpose of Lexical Markup Framework (LMF, ISO-24613) is to define a standard for lexicons that covers multilingual lexical data.
This paper is the description of the ongoing work within ISO committees and is not a position paper.

KEY WORDS: LMF standardization lexicon multilingual ISO-TC37

1 Introduction

Lexical Markup Framework (LMF) is a model that provides a common standardized framework for the construction of Natural Language Processing (NLP) lexicons. The goals of LMF are to provide a common model for the creation and use of lexical resources, to manage the exchange of data between and among these resources, and to enable the merging of a large number of individual electronic resources to form extensive global electronic resources. The descriptions addressed by the standard proposal range from morphology, syntax and semantics to translation information organized as different extensions of an obligatory descriptive core package. LMF is intended for NLP lexicons to be used in a non-restricted range of applications and lan-

guages. LMF is also intended for machine readable dictionaries (MRD), which are not within the scope of this paper.
In this paper, we offer a snapshot of how the standard proposed for representing multilingual information looks like. The full technical specification may be found in [LMF 2008].

2 History and current context

In the past, the standardization of the formal description of lexical resources has been studied and addressed by a series of projects like EUROTRA-7, GENELEX [Antoni-Lay 1994], MULTEXT [Ide 1994], EAGLES [Calzolari 1996], PAROLE [Zampolli 1997], SIMPLE [Lenci 2000], ISLE [Atkins 2002] and MILE [Bertagna 2004]. Although the standards issued by these projects had been widely adopted by research institutions and academy, they also needed adoption within the industrial community to support advanced language technologies for content access and sharing. In order to reach wide industrial audience, production and ratification by an official International body seemed necessary. In 2002, the ISO-TC37 National delegations decided to address standards dedicated to resources for NLP.

These standards are currently elaborated as high level specifications and deal with word segmentation (ISO 24614), annotations (ISO 24611, 24612 and 24615), feature structures (ISO 24610), and lexicons (ISO 24613), with this latest one being the focus of the current paper. ISO 24613 or LMF owes the past for the major standardization activities and *best-practices* in the field it is actually built upon.

These standards deploy low level specifications dedicated to constants, namely data categories (revision of ISO 12620), language codes (ISO 639 or IETF BCP-47 *tags for the identification of languages*), script codes (ISO 15924), country codes (ISO 3166), dates (ISO 8601) and Unicode (ISO 10646).

This is the essence of the “structure-adornment” binomial which neatly separates the standardization effort into high-level specification (the structure) and low-level specification (the adornment). In LMF, this combination allows the implementation of standard-conformant lexical resources.

The two level organization has been devised to form a coherent family of standards with the following simple rules:
1) the **high level specifications** provide structural classes. Each class is defined by a name, an English text describing its usage and a formal specification of the relations with the other classes. These structural classes are intended to be adorned by constants and attributes.
2) the **low level specifications** provide standardized constants and attribute name.

3 Scope and challenges

The aim of LMF efforts is directed to elaborate a proposal that tries to face the challenges posed by most of existing lexical models which are complex and very different in nature from each other, because they contain different types of information.

LMF addresses the following topics:

- Represent words in languages where multiple orthographies (native scripts or transliterations) are possible, e.g. some Asian languages.
- Represent explicitly (i.e. in extension) the morphology of languages where a description of all inflected forms (from a list of lemmatised forms) is manageable, e.g. English.
- Represent the morphology of languages where a description in extension of all inflected forms is not manageable (e.g. Hungarian). In this case, representation in intension is the only manageable way and a mechanism called "morphological pattern" is provided for this purpose.
- Easily associate written forms and spoken forms for all languages.
- Represent complex agglutinating compound words like in German.
- Represent fixed, semi-fixed and flexible multiword expressions.
- Represent specific syntactic behaviors, as defined by EAGLES.
- Allow complex argument mapping between syntactic and semantic descriptions, as defined by EAGLES.
- Allow a semantic organisation based on SynSets (like in WordNet) or on semantic predicates (like in FrameNet).
- Represent large scale multilingual resources based on interlingual pivots or on transfer linking.

4 Modeling standard used by LMF

The LMF specification complies with the modeling principles of Unified Modeling Language (UML) as defined by the Object Management Group (OMG) [Rumbaugh 2004]. UML is a general-purpose visual modeling language that is used to specify, visualize, construct and document data structures. The modeling language is intended to unify past experience of modeling techniques and to incorporate current software best practices into a coherent approach.

UML has been chosen for the following reasons:

- UML is the 'de facto' standard for modeling in the Industry. That means that a lot of professionals are able to understand the specifications.
- UML is well defined and documented;
- the use of diagrams is very efficient when a model needs to be presented and negotiated[1]. It is a perfect language for modeling and has a very large and rapidly expanding user community. With respect to other representation languages, UML allows to work at different layers of abstraction, zooming out from a detailed view to the overall environment and is particularly suited to human users;
- UML allows designers (and readers) to partition large models into workable pieces by means of UML packages;
- Various powerful UML tools are available now in order to ease the design process.

UML captures information about the static structure and dynamic behavior of a system, but in LMF, we restrict ourselves to the static aspect. We also provide informative examples of content markup using another key standard, XML, although XML is just one way of expressing a LMF model. We defined an XML DTD for the purpose of driving any LMF process and designing concrete lexicon instances. This DTD can be used automatically by a program to check the conformance of a given lexicon.

5 Structure and core package

LMF sticks to the very well consolidated ISO strategy to split the specification into two separate objects: the structure and the content. LMF defines the structure of the lexicon while the features that encode information in form of attribute-value pairs are not defined here but are recorded in the ISO Data Category Registry as specified by ISO-12620. More precisely, LMF defines class names, class usages, class relations by means of English texts and UML diagrams. This specification goes with some guidelines and a series of examples, but it is important to highlight that attribute-value pairs like /grammatical gender/ and /feminine/ are not defined within LMF.

LMF is comprised of two types of packages:
1) the **core package** that consists of a structural skeleton in order to represent the basic hierarchy of information in a lexicon.
2) **extensions to the core package** that reuse the core classes in conjunction with additional classes required for the description of the contents of a specific lexical resource.

The core package is specified by the following UML class model:

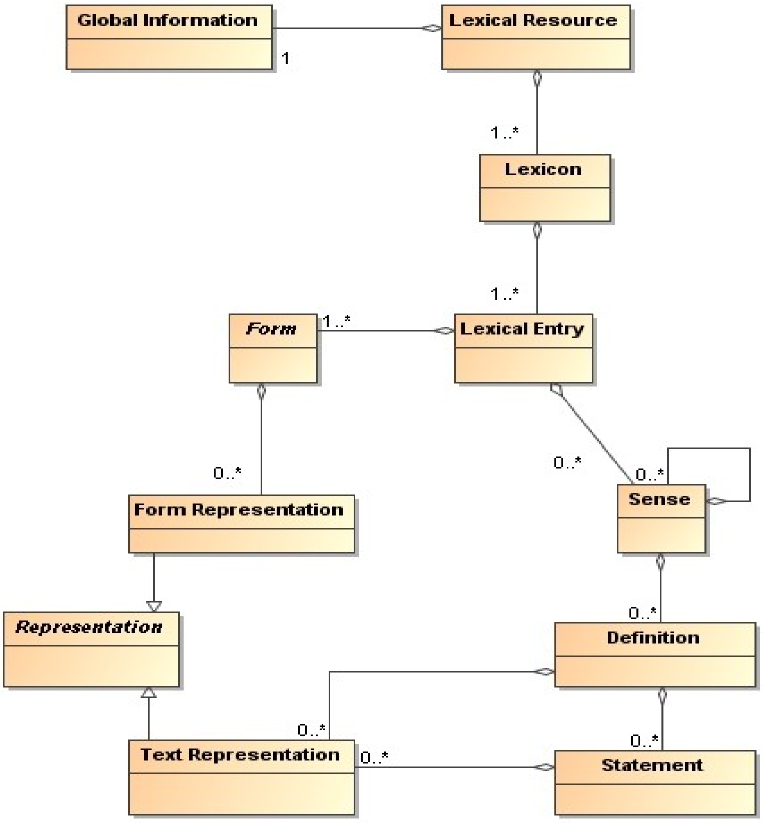

Figure 1: core model

The class called *Lexical Resource* represents the entire resource and is a container for one or more lexicons. The *Global Information* class contains administrative information and other general attributes. The *Lexicon* class is the container for all the lexical entries of the same language.

The *Lexical Entry* class is a container for managing the top level language instances. As a consequence, the number of representatives of single words, multi-word expressions and affixes of the lexicon is equal to the number of lexical entries in a given lexicon. The *Form* and *Sense* classes are parts of the *Lexical Entry*. Therefore, the *Lexical Entry* manages the relationship between sets of related forms and their senses.

If there is more than one orthography for the word form (e.g. transliteration) the *Form* class may be associated with one to many *Form Representations*, each of which contains a specific orthography and one to many data categories that describe the attributes of that orthography.

Definition is a class representing a narrative description of a sense. It is displayed for human users to facilitate their understanding of a *Lexical Entry* and is not meant to be processable by computer programs. Each *Definition* instance may be associated with zero to many *Text Representation* instances in order to manage the text definition in more than one language or script. *Statement* is a class representing a narrative description and refines or complements *Definition*.

From the point of view of UML, an extension is a UML package. Current extensions for NLP dictionaries are: NLP Morphology[2], NLP Morphological pattern, NLP Multiword expression pattern, NLP Syntax, NLP Semantic, Constraint expression and Multilingual notations, which is the focus of this paper.

6 NLP Multilingual notation Extension

6.1 Overview

The NLP multilingual notation extension is dedicated to the description of the mapping between two or more languages in a LMF resource. The model is based on the notion of *Axis* that links *Senses*, *Syntactic Behavior* and *Context* that are defined in semantic, syntactic, and MRD packages. *Syntactic Behaviour* is a class representing one of the possible behaviours of a word. *Context* is a class used to illustrate the particular meaning of a *Sense* instance. *Axis* is a term taken from the Papillon[3] project [Sérasset 2001][4]. *Axis* instances can be organized at the lexicon manager convenience in order to link directly or indirectly objects of different languages.

6.2 Considerations for standardizing multilingual data

The simplest configuration of multilingual data is a bilingual lexicon where a single link is used to represent the translation of a given form/ sense pair from one language into another. But a survey of actual practices clearly reveals other requirements that make the model more complex.

Consequently, LMF has focused on the following ones:

1) Cases where the relation 1-to-1 is impossible because of lexical differences among languages. This is usually called diversification and neutralization. An example is the English word "river" that relates to French words "rivière" and "fleuve", where the latter is used for specifying that the referent is a river that flows into the sea. The bilingual lexicon should specify how these units relate.

2) The bilingual lexicon approach should be optimized to allow the easiest management of large databases for real multilingual scenarios. In order to reduce the explosion of links in a multi-bilingual scenario, translation equivalence can be managed through an intermediate *Axis*. This object can be shared in order to contain the number of links in manageable proportions.

3) The model should cover both *transfer* and *pivot* approaches to translation, taking also into account hybrid approaches. In LMF, the pivot approach is implemented by a *Sense Axis*. The transfer approach is implemented by a *Transfer Axis*.

4) A situation that is not very easy to deal with is how to represent translations to languages that are similar or variants. The problem arises, for instance, when the task is to represent translations from English to both European Portuguese and Brazilian Portuguese. It is difficult to consider them as two separate languages. In fact, one is a variant of the other. The differences are minor: a certain number of

words are different and some limited phenomena in syntax are different. Instead of managing two distinct copies, it is more effective to manage one lexicon with some objects that are marked with a dialectal attribute. Concerning the translation from English to Portuguese: a limited number of specific Axis instances record this variation and the vast majority of Axis instances is shared.

5) The model should allow for representing the information that restricts or conditions the translations. The representation of tests that combine logical operations upon syntactic and semantic features must be covered.

6.3 Structure

The model is based on the notion of *Axis* that link *Senses*, *Syntactic Behavior* and *Context* instances pertaining to different languages. An *Axis* instance is not specific to a given language: its scope is the whole database, thus, *Axis* instances are not aggregated in a *Lexicon* instance like lexical entries but are aggregated in the *Lexical Resource* instance.

Axis instances can be organized at the lexicon manager convenience in order to link directly or indirectly objects of different languages. A direct link is implemented by a single axis. An indirect link is implemented by several axis and one or several relations.

The UML class model is a UML package as follows:

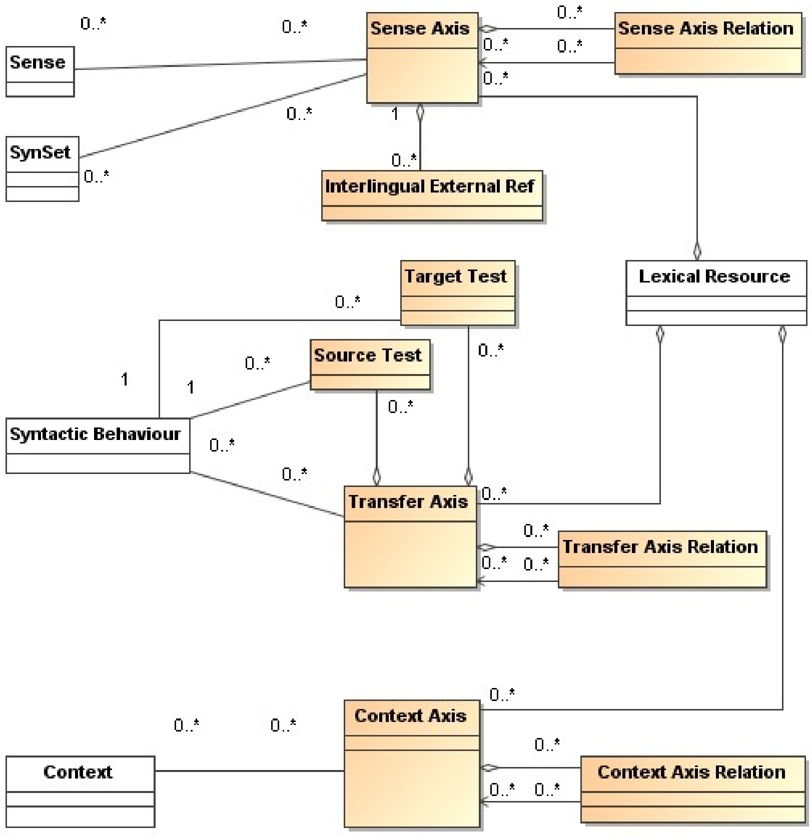

Figure 2: multilingual notations model

6.4 Sense Axis

Sense Axis is used to link closely related senses in different languages, under the same assumptions of the interlingual pivot approach. The use of the *Sense Axis* facilitates the representation of the translation of words that do not necessarily have the same valence or morphological form in one language than in another.

6.5 Interlingual External Ref

A *Sense Axis* instance may be associated with one or several *Interlingual External Ref* instances. It is not the purpose of this model to code a complex system for knowledge representation, which ideally should be structured as a complete coherent system designed specifically for this purpose. But it may be useful to define a bridge to one or several system*s. Interlingual External Ref* is provided for this particular purpose.

6.6 Sense Axis Relation

*Sense Axis Relatio*n permits to describe the linking between two different *Sense Axis* instances. The label enables the coding of simple interlingual relations like the specialization of “fleuve” compared to “rivière” and “river”.

6.7 Transfer Axis

Transfer Axis is designed to represent multilingual transfer approach. Here, linkage refers to information contained in syntax. For example, this approach enables the representation of syntactic actants involving inversion, such as : fra:“elle me manque” => eng:“I miss her”.

6.8 Transfer Axis Relation

Transfer Axis Relation links two *Transfer Axis* instances.

6.9 Source Test and Target Test

Source Test permits to express a condition on the translation on the source language side while *Target Test* does it on the target language side.

6.10 Context Axis

Context Axis supplies documentation for sample translations. The purpose is not to record large scale multilingual corpora. The goal is to link a Lexical Entry with a typical example of translation.

6.11 Context Axis Relation

Context Axis Relation links *Context Axis* instances.

7 Two examples

7.1 Simple example of a near match

The first example is about the interlingual approach with two axis instances to represent a near match between "fleuve" in French and "river" in English. There are two senses in French and one sense in English. In the diagram, French is located on the left side and English on the right side. Multilingual notations are located in the middle. The axis on the top implements a direct semantic equivalence between the two languages for the relation that holds between "rivière" and "river". But, while there is a semantic relation between the two French senses, the axis of the more specific term in French is not linked directly to any English sense because this notion does not exist in English.

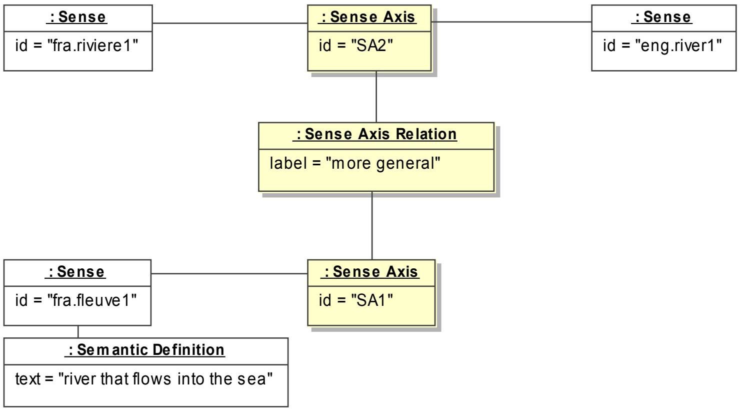

Figure 3: simple example of a near match

7.2 Example in three languages of a shared transfer structure

A second example shows how to use the Transfer Axis relation to relate different information in a multilingual transfer lexicon.

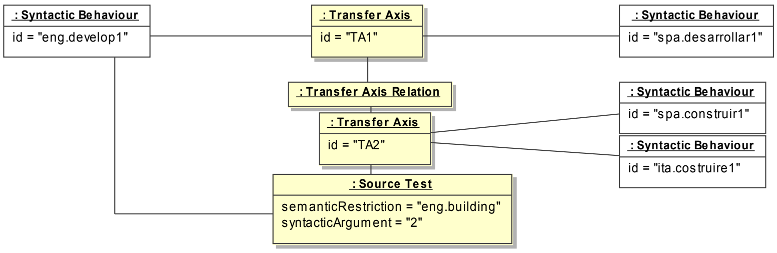

Figure 4: example in three languages of a shared transfer structure

It represents the translation of the English “develop” into Italian and Spanish. While the more general sense links the English “develop” and the Spanish “desarrollar”, a second correspondence expresses restrictions that should be tested in the source language: if the second argument of the construction refers to a certain element (for instance, a building) it should be translated into specific Spanish or Italian verbs.

8 Other modeling options for multilingual notations

Other lexical structures have been studied and discussed during the numerous ISO meetings and email exchanges. For NLP lexicons, we did not retain models based on simple bilingual links because when the number of pairs of language increases, the number of links explodes to unmanageable proportions. Such an organization cannot be named as multilingual. Another option would have been to consider that the notion of a concept is the most important notion in the resource. According to this organization (usually named as onomasiological) the data are structured as a set of trees (a forest) that are aggregated within a global resource. The concepts are the roots and the lexical written forms are the leaves. This organization is the one retained by TBX (i.e. ISO DIS-30042) for instance. This structure is quite simple and is well suited for simple mono and/or multilingual terminologies but it appears that language representations do not fit well within such an organization because most relations are transverse. The main point is that linguistic descriptions, for a given language need to be both more powerful and highly shared.
More precisely:

- Syntactic representations like subcategorization frames need more complex structures (possibly recursive) that require to be highly shared among certain classes of words.
- Transfer representations (see *Transfer Axis* class in LMF) that are so important for machine translation do not fit within this organization because they are transverse.
- An interlingual pivot must not be mandatory for the words that are specific to a given language or culture. This situation appears for proper nouns like "NBA" for instance. A good lexical model must allow the lexicon manager to keep local to a language what is considered as local.
- Morphological patterns that are mandatory for representing complex languages like Hungarian or Arabic must be defined and shared.
- Multiword expression patterns must also be defined and shared.

The option that we retained is to have both the notion of lexicon (holding language specific representations) and the notion lexical resource (holding interlingual axes). This is more powerful and flexible. Nevertheless, if a user wants to have only interlingual axes, LMF allows this option. This user just have to use the notion of lexical resource and to manage *Sense Axis* instances. But obviously, as a consequence, such an NLP lexicon without any morphology or syntax does not allow very powerful processings.

9 LMF in XML

9.1 Chosen option

A DTD is provided as an informative annex in the ISO document [LMF 2008]. Based on this DTD, the first example (i.e. "river") can be serialized with the following XML tags:

```
<LexicalResource>
        <GlobalInformation>
                <feat att="languageCoding" val="ISO 639-3"/>
        </GlobalInformation>
  <!—                              French section  -->
  <Lexicon>
        <feat att="name" val=”French Extract”/>
        <feat att="language" val="fra"/>
    <LexicalEntry >
        <feat att="partOfSpeech" val=”noun”/>
        <Lemma>
                <feat att="wordForm" val=”fleuve”/>
        </Lemma>
        <Sense id=”fra.fleuve1”>
              <SemanticDefinition>
                <feat att="text" val=”Grande rivière lorsqu'elle aboutit à la mer”/>
                <feat att="source" val=”Le Petit Robert 2003”/>
                </SemanticDefinition>
        </Sense>
    </LexicalEntry>
    <LexicalEntry>
        <feat att="partOfSpeech" val=”noun”/>
        <Lemma>
              <feat att="wordForm" val=”rivière”/>
        </Lemma>
        <Sense id=”fra.riviere1”>
              <SemanticDefinition>
                <feat att="text" val=”Cours d'eau naturel de moyenne importance”/>
                <feat att="source" val=”Le Petit Robert 2003”/>
                </SemanticDefinition>
        </Sense>
    </LexicalEntry>
  </Lexicon>
<!—                                   Multilingual section -->
  <SenseAxis id=”A1” senses="fra.fleuve1">
        <SenseAxisRelation targets="A2">
              <feat att="comment" val="flows into the sea"/>
```

```
            <feat att="label" val="more precise"/>
        </SenseAxisRelation>
  </SenseAxis>
  <SenseAxis id=”A2” senses="fra.riviere1 eng.river1"/>
  <!—                                    English section -->
  <Lexicon>
        <feat att="name" val=”English Extract”/>
        <feat att="language" val="eng"/>
    <LexicalEntry>
        <feat att="partOfSpeech" val=”noun”/>
        <Lemma>
              <feat att="wordForm" val=”river”/>
        </Lemma>
        <Sense id=”eng.river1”>
             <SemanticDefinition>
              <feat att="text" val=”A large permanent body of flowing water, originating at a
source, travelling along a fixed course, and emptying into a lake or the sea”/>
              <feat att="source" val=”Harraps Chambers 2005”/>
              </SemanticDefinition>
        </Sense>
    </LexicalEntry>
  </Lexicon>
</LexicalResource>
```

9.2 Other options

There might be differing modeling approaches concerning XML serialization. We decided to use an XML DTD in the ISO document because :

- a DTD is still the most accessible mechanism for tag structure, due to the fact that our users are not necessary experts in XML;
- a DTD is concise thus is more easily readable than a more lengthy specification.

But three other technical options are possible: a Relax-NG specification (i.e. ISO 19757-2), a W3C schema [5] or a RDF description. In the near future, if our users require such specifications we could easily write an additional ISO technical report that will be associated to the LMF document.

10 Comparison

A serious comparison with previously existing models and concrete usage of LMF is not possible in this current paper due to the lack of space. We advice the interested colleague to consult the technical report "Extended examples of lexicons using LMF" located at: "http://lirics.loria.fr" in the document area ; see also [Khemakhem et al 2007].

11 Conclusion

Currently (Spring 2008), LMF is in *Final Draft for International Standard* (DIS) stage. We schedule to reach final *International Standard* (IS) stage in Winter 2008[6].

In this paper we presented the results of the ongoing research activity of the LMF ISO standard. In order to reach a consensus, the work done has paid careful attention to the similarities and differences of existing lexicons and the models behind them. In the future, the LMF users will be able to:

- use an interoperable model;
- have a model that allows a wide range of representations;
- use standard based tools like interactive software platforms, lexicon mergers or web services access.

Acknowledgements The work presented here was partially funded by the EU eContent-22236 LIRICS project[7].

References

Antoni-Lay M-H., Francopoulo G., Zaysser L. 1994 A generic model for reusable lexicons: the GENELEX project. Literary and linguistic computing 9(1)

Atkins S., Bel N., Bertagna F., Bouillon P., Calzolari N., Fellbaum C., Grishman R., Lenci A., MacLeod C., Palmer M., Thurmair G.,Villegas M., Zampolli A., 2002. From Resources to Applications. Designing the Multilingual ISLE Lexical Entry, Proceedings of LREC Las Palmas

Bertagna F., Lenci A., Monachini M., Calzolari N. 2004 Content interoperability of lexical resources, open issues and MILE perspectives, Proceedings of LREC Lisbon

Calzolari N., Mc Naught J., Zampolli A. 1996 Eagles, editors introduction www.ilc.cnr.it/EAGLES96/edintro.html

Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. 2006 Lexical Markup Framework (LMF), Proceedings of LREC Genoa

Ide N., Véronis J., 1994. MULTEXT: Multilingual Text Tools and Corpora. In Proceedings of the 15th International Conference on Computational Linguistics, COLING Kyoto

Khemakhem A., Gargouri B., Abdelwahed A., Francopoulo G. 2007. Modélisation des paradigmes de flexion des verbes arabes selon la norme LMF-ISO 24613. TALN Toulouse

Lenci A., Bel N., Busa F., Calzolari N., Gola E., Monachini M., Ogonowski A., Peters I., Peters W., Ruimy N., Villegas M., Zampolli A. 2000. SIMPLE: A General Framework for the Development of Multilingual Lexicons, International Journal of Lexicography 13(4). Oxford University

LMF 2008 Lexical Markup Framework ISO-FDIS24613, ISO Geneva

Rumbaugh J., Jacobson I., Booch G. 2004 The unified modeling language reference manual, second edition, Addison Wesley

Sérasset G., Mangeot-Lerebours M. 2001 Papillon Lexical Database project: monolingual dictionaries & interlingual links NLPRS Tokyo

Van der Vlist E. 2004 Relax NG, O'Reilly, Sebastopol, California

Zampolli A. 1997 The PAROLE project in the general context of the European actions for Language Resources. In: Marcinkeviciene, R., Volz, N. (eds.): Telri Proceedings of the Second European Seminar: Language Applications for a Multilingual Europe. IDS/VDU, Manheim/Kaunas.

[1] Two types of diagrams must be distinguished: class diagrams and instance diagrams. A model is specified by a UML class diagram within a UML package: in this case, the class name is not underlined in the diagram. The various examples of word description are represented by UML instance diagrams: in this case, the class name is underlined.

[2] Packages for morphology, syntax and Semantics are described in [Francopoulo 2006].

[3] www.papillon-dictionary.org

[4] To be more precise, Papillon uses the term "axie" from "axis" and "lexie". In the beginning of the LMF project, we used the term "axie" but after some bad comments about using a non-English term in a standard, we decided to use the term "axis".

[5] It should be noted that a W3C schema is not specified as an ISO standard but is specified as a W3C recommandation. For criticisms about W3C schemas and comparison with Relax NG, see [Van der Vlist 2004].

[6] Please consult www.lexicalmarkupframework.org for updated information.

[7] http://lirics.loria.fr